Molecular Ecology Resources (2008) **8**, 765–768 doi: 10.1111/j.1755-0998.2008.02128.x

COMPUTER PROGRAMS

P-LOCI: a computer program for choosing the most efficient set of loci for parentage assignment

SEAN E. MATSON,* MARK D. CAMARA,† WILL EICHERT‡ and MICHAEL A. BANKS‡
*Department of Animal Sciences, Oregon State University, 2030 SE Marine Science Drive, Newport, OR 97365-5229, USA,
†Shellfish Genetics Program, USDA Agricultural Research Service, 2030 SE Marine Science Drive, Newport, OR 97365-5229, USA
‡Department of Fisheries and Wildlife, Coastal Oregon Marine Experiment Station, Oregon State University, 2030 SE Marine Science Drive, Newport, OR 97365-5229, USA

Abstract

Determining how many and which codominant marker loci are required for accurate parentage assignment is not straightforward because levels of marker polymorphism, linkage, allelic distributions among potential parents and other factors produce differences in the discriminatory power of individual markers and sets of markers. P-LOCI software identifies the most efficient set of codominant markers for assigning parentage at a user-defined level of success, using either simulated or actual offspring genotypes of known parentage. Simulations can incorporate linkage among markers, mating design and frequencies of null alleles and/or genotyping errors. P-LOCI is available for windows systems at http://marineresearch.oregonstate.edu/genetics/ploci.htm.

Keywords: codominant markers, linkage, microsatellite, offspring simulation, parentage assignment, SNP

Received 23 September 2007; revision accepted 13 January 2008

Parentage assignment using codominant molecular markers has become increasingly important for quantitative genetics, animal breeding, molecular ecology and evolutionary biology (Vignal *et al*. 2002; Jones & Ardren 2003; Anderson & Garza 2006). Determining the most efficient set of marker loci to use for a particular set of parents can save considerable time, effort and funds. The minimum number of loci necessary to accurately assign parentage depends on a number of factors that affect their informativeness, including allelic richness and diversity, linkage disequilibrium among marker loci due to physical linkage and other sources, number of parental pairs, mating design, frequency of null alleles and genotyping errors, and unequal numbers of offspring per family (Bernatchez & Duchesne 2000; Jones & Ardren 2003; Dakin & Avise 2004; Anderson & Garza 2006; Kalinowski & Taper 2007; Kalinowski *et al*. 2007). Few currently available parentage software packages have multilocus predictive capabilities, and they do not incorporate many of these important factors (Jones & Ardren 2003; Taggart 2007). Most researchers and all currently available parentage software assume markers are not linked, even though physically linked markers carry redundant information and are thus less informative in combination than expected from single locus characteristics. P-LOCI is the only program that uses linkage information together with variable locus-specific frequencies of null alleles and genotyping errors in the simulation of offspring genotypes with variable number of offspring per family to determine the minimum set of loci for assigning parentage. Additionally, because the best combination of loci can vary among populations, marker informativeness must be re-evaluated for each study population, creating the need for a quick and easy to use software tool. We created P-LOCI to increase the efficiency of parentage assignment by quickly identifying the best available set of codominant molecular markers for parentage assignment in a specific population. Figure 1 shows the P-LOCI interface with an explanation of the controls.

P-LOCI identifies the smallest suite of codominant loci required to assign diploid offspring to their parents at a user-defined level of success through an iterative procedure. In either simulation or real progeny mode, the user provides a parental file consisting of the candidate parents' multilocus

Correspondence: Sean E. Matson, Fax: 541-867-0105; E-mail: matsonse@onid.orst.edu

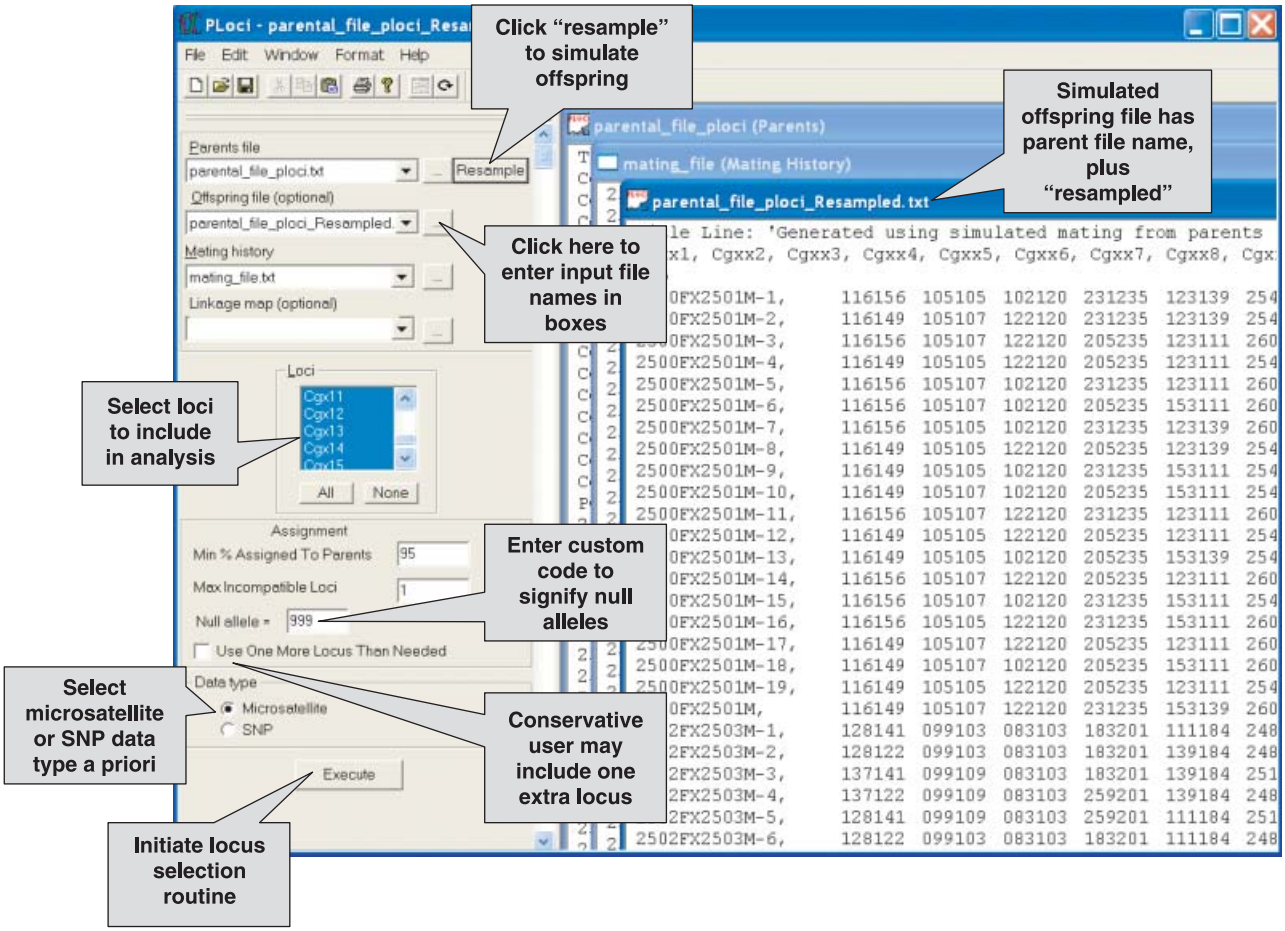

Fig. 1 Screenshot of P-LOCI software interface, showing where to enter input files and other important operating information.

marker genotypes at all loci to be evaluated and a mating design file specifying how the parents are paired. When the mating structure is not known, the user submits an all-combination mating file. P-LOCI simulates offspring genotypes using those files and optional linkage and error information, and then attempts to assign them to their parents based on using an exclusion algorithm (Jones & Arden 2003). The accuracy of these assignments is evaluated against the known pedigrees of simulated or actual progeny. P-LOCI was created primarily for use with microsatellite data, but works with any codominant genetic markers. Figure 2 is a conceptual model of P-LOCI, showing information flow.

P-LOCI simulates biologically realistic offspring genotypes through a computationally intensive but genetically realistic 'brute force' procedure by first building virtual gametic haplotypes from each parent. For each virtual offspring, the program first randomly chooses one allele from the current parent at the first locus in each linkage group and then 'walks' along the virtual parental meiotic chromatid. Cross-over probabilities between adjacent markers are determined by recombination fractions calculated from linkage map distances. If a linkage map is not provided, the program assumes independent marker segregation and assembles each haplotype choosing each allele at each locus with equal probability. The two haplotypes are then combined into a diploid offspring according to the mating design.

P-LOCI accommodates different male and female maps in either Kosambi or Haldane distances (Liu 1997; Lynch & Walsh 1998). Linkage phase among marker alleles in specific parents is assumed as their order of entry in the parental genotype file. If the user knows the phase, they can enter it as such, although the true phase is usually unknown, and therefore arbitrarily represented in genotype data. The user can also vary the number of offspring produced per family. This may be desirable if some families are expected to be over-represented in the offspring pool, or to model variability in the best marker set, due to variance in relative contribution of specific parents to the offspring population.

P-LOCI can realistically incorporate two types of error when simulating offspring genotypes: segregating null alleles and random-genotyping errors. P-LOCI optionally introduces

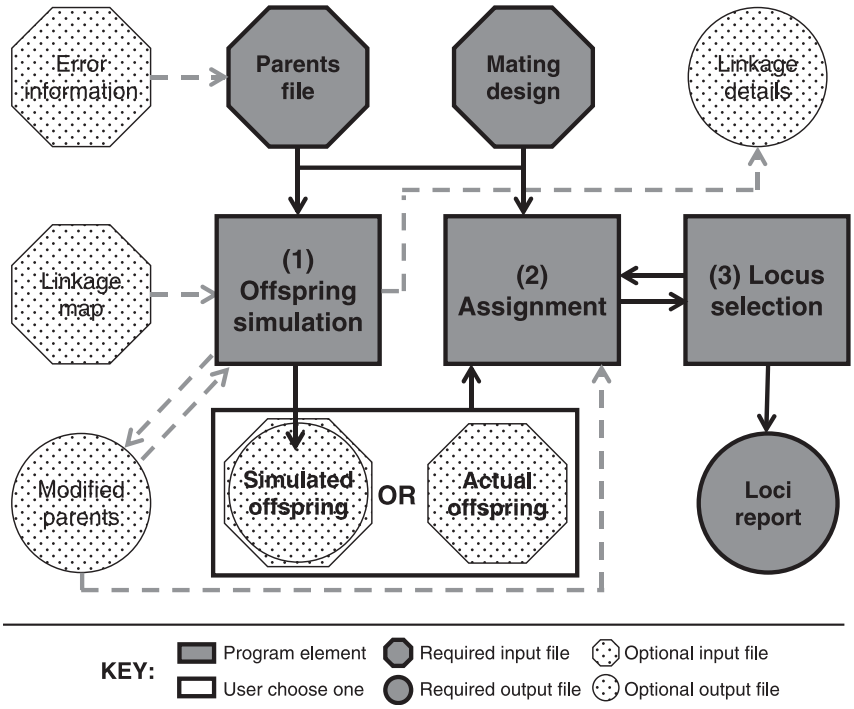

Fig. 2 Conceptual model of P-LOCI showing information flow, signified by arrow direction. Program operation is depicted as follows: (1) P-LOCI simulates offspring (or they are provided by the user); (2) Those offspring are assigned to candidate parental pairs, as denoted in the mating design; (3) The locus selection routine ranks locus sets by assignment success, and either accepts the locus set as fulfilling the assignment success criterion, or reinitiates the assignment routine to perform all possible combinations of the next level (e.g. from pairs to triplets), or the current locus combination fulfills the criterion, the program stops and reports.

null alleles at user-specified frequencies and creates a modified parental genotype file in which a proportion of the homozygous parents at each locus are re-coded as heterozygotes with an undetectable null allele. The simulated offspring that inherit these null alleles are treated by the assignment algorithm as homozygotes for the detectable allele. Null allele frequencies and genotyping error rates must be estimated by the user a priori using other available software (e.g. van Oosterhout *et al*. 2004; Kalinowski *et al*. 2007). The user can also designate individual parents a priori as null homozygotes.

P-LOCI optionally incorporates microsatellite marker typing errors by randomly adding or subtracting a user-defined number of base pairs to the offspring alleles, producing mismatches and potential misassignments that realistically compromise the discriminatory value of error-prone loci. To mitigate errors in real or simulated data sets that prevent assignment via exclusion, the user can enter a maximum number of loci at which offspring are allowed to mismatch potential parents and still be assigned to them. The conservative user can also have P-LOCI determine a marker set with one more locus than is needed to reach the assignment success criterion.

If information regarding the rates of null alleles, typing errors or linkage relationships among markers is not available, the user may wish to genotype a small number of offspring of known parentage from all crosses (e.g. offspring of controlled crosses or observed matings) at all loci in order to produce an input file containing actual offspring genotypes rather than simulated ones. Actual offspring genotypes will inherently exhibit the effects of the aforementioned complicating factors.

After P-LOCI either produces the simulated offspring file or is provided with actual offspring genotypes of known parentage, the user initiates the marker evaluation algorithm, and P-LOCI first assigns all offspring using each marker individually by checking each offspring for Mendelian compatibility with each parental pair in the mating file. Assignments are successful only when a single compatible parental pair is identified. If more than one compatible pair is found, or if an offspring is misassigned when checked against its known parentage information, that individual assignment is unsuccessful. This information is used to rank individual loci by their assignment success rate. The software subsequently examines all possible marker pairs, triplets, etc., and stops when it reaches the user-provided level of assignment success. The program then produces a report that includes the ranks of individual loci and their assignment scores, followed by the best pair, triplet and so on. P-LOCI can automatically produce and analyse multiple sets of simulated offspring and produce a summary report that includes the average rankings of individual loci among runs and how often a particular locus appeared in the best marker set. After using P-LOCI to determine the best set of loci, the user can assign actual progeny to their parents using a variety of methods and software, of which Jones & Ardren (2003) provide a thorough review.

We tested P-LOCI with actual and simulated microsatellite and single nucleotide polymorphism (SNP) data, varying levels of polymorphism, distribution of alleles among parents, number of parents, mating design complexity, degree of linkage among markers, and locus-specific frequencies of null alleles and genotyping errors. P-LOCI reported increasing assignment success with increasing allelic richness and more heterogeneous allelic distributions among potential parents. Assignment success decreased with increasing number of potential parents, increasing complexity of the mating design, and higher frequencies of null alleles and genotyping errors. In general, unlinked markers provided a higher level of assignment success than linked ones, all other factors being equal.

P-LOCI also chose different marker sets in different parental populations of Pacific oysters, *Crassostrea gigas* (S.E. Matson and M.D. Camara, C. Langdon and F. Evans, unpublished data), using genotype data from breeding experiments, microsatellite markers, and microsatellite linkage map information (Hubert & Hedgecock 2004). We used an early version of P-LOCI to determine the best available suite of microsatellites (Li *et al*. 2003; Magoulas *et al*. 1998; McGoldrick *et al*. 2000) for assigning 1200 offspring to 20 pairs of parents, and performed parentage analysis with a 98.5% success rate using PAPA software (Duchesne *et al*. 2002) and four loci.

Single locus ranks within populations were similar to those obtained from ranking loci by Shannon diversity index computed with MICROSATELLITE ANALYSER (Dieringer & Schlotterer 2003). However, we found that the best suite of loci often consists not of only the top-ranked individual loci, but rather a mixture of top- and middle-ranked markers. This is most likely due to random allelic associations among loci, and LD in the parental population that make the information carried by some marker sets redundant and others complementary.

Our preliminary results have important implications. The top ranked individual loci do not necessarily constitute the smallest group of loci for assignment, and that group is not necessarily the best for all populations, making P-LOCI an important tool for efficient parentage analysis.

Acknowledgements

This research was funded by Oregon Sea Grant, the California Department of Water Resources (Environmental Services), the USDA Agricultural Research Service (CRIS Project #5358-31000-001-00D), and a Mamie Markham Research Award at OSU-HMSC. We thank Chris Langdon and Ford Evans for use of their oyster genotype data.

References

Anderson EC, Garza JC (2006) The power of single-nucleotide polymorphisms for parentage inference. *Genetics*, **172**, 2567–2582.

Bernatchez L, Duchesne P (2000) Individual-based genotype analysis in studies of parentage and population assignment: how many loci, how many alleles? *Canadian Journal of Fisheries and Aquatic Sciences*, **57**, 1–12.

Dakin EE, Avise JC (2004) Microsatellite null alleles in parentage analysis. *Heredity*, **93**, 504–509.

Dieringer D, Schlotterer C (2003) MICROSATELLITE ANALYSER (MSA): a platform independent analysis tool for large microsatellite data sets. *Molecular Ecology Notes*, **3**, 167–169.

Duchesne P, Godbout MH, Bernatchez L (2002) PAPA (package for the analysis of parental allocation): a computer program for simulated and real parental allocation. *Molecular Ecology Notes*, **2**, 191–193.

Hubert S, Hedgecock D (2004) Linkage maps of microsatellite DNA markers for the Pacific oyster, *Crassostrea gigas*. *Genetics*, **168**, 351–362.

Jones AG, Ardren WR (2003) Methods of parentage analysis in natural populations. *Molecular Ecology*, **12**, 2511–2523.

Kalinowski ST, Taper ML (2007) Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. *Conservation Genetics*, **7**, 991–995.

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, **16**, 1099–1106.

Li G, Hubert S, Bucklin K, Ribes V, Hedgecock D (2003) Characterization of 79 microsatellite DNA markers in the Pacific oyster *Crassostrea gigas*. *Molecular Ecology Notes*, **3**, 228–232.

Liu B (1997) *Statistical Genomics: Linkage, Mapping and QTL Analysis*. CRC Press LLC, Boca Raton, Florida.

Lynch M, Walsh B (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer & Associates Inc., Sunderland, Massachusetts.

Magoulas A, Gjetvaj B, Terzoglou V, Zouros E (1998) Three polymorphic microsatellites in the Japanese oyster, *Crassostrea gigas* (Thunberg). *Animal Genetics*, **29**, 69–70.

McGoldrick DJ, Hedgecock D, English LJ, Baoprasertkul P, Ward RD (2000) The transmission of microsatellite alleles in Australian and North American stocks of the Pacific oyster (*Crassostrea gigas*): selection and null alleles. *Journal of Shellfish Research*, **19**, 779–788.

Taggart JB (2007) FAP: an exclusion-based parental assignment program with enhanced predictive functions. *Molecular Ecology Notes*, **7**, 412–415.

van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes*, **4**, 535–538.

Vignal A, Milan D, SanCristobal M, Eggen A (2002) A review on SNP and other types of molecular markers and their use in animal genetics. *Genetic Selection and Evolution*, **34**, 275–305.